Abstract—Principal Component Analysis (PCA) is one of the most widely used subspace projection technique for face recognition. In subspace methods like PCA, feature selection is fundamental to obtain better face recognition. However, the problem of finding a subset of features from a high dimensional feature set is $NP$-hard. Therefore, to solve the feature selection problem, heuristic methods such as evolutionary algorithms are gaining importance. In many face recognition applications, due to the small sample size (SSS) problem, it is difficult to construct a single strong classifier. Recently, ensemble learning in face recognition is gaining significance due to its ability to overcome the SSS problem. In this paper, the $NP$-hard problem of finding the best subset of the extracted PCA features for face recognition is solved by using the differential evolution (DE) algorithm and is referred to as FS-DE. The feature subset is obtained by maximizing the class separation in the training data. We also present an ensemble based approach for face recognition (En-FR), where different subsets of PCA features are obtained by maximizing the distance between a subset of classes of the training data instead of whole classes. The subsets of the classes are obtained by bagging and overlap each other. Each subset of the PCA features selected is used for face recognition and all the outputs are combined by a simple majority voting. The proposed algorithms, FS-DE and En-FR, are evaluated on four well-known face databases and the performance is compared with the PCA and Fisher’s LDA algorithms.

Keywords—face recognition (FR); principal component analysis (PCA); ensemble learning; differential evolution; machine learning; small-sample-size (SSS) problem; feature selection.

I. INTRODUCTION

Face recognition in unconstrained environments is of great interest and challenge in the area of computer vision. Face recognition algorithms that utilize the intensity or intensity-derived features of the original images suffer from the “curse of dimensionality” even for images of modest size. As the dimensionality of the input images increases, the dimensionality of the feature vectors increases. In applications like face recognition, where the training sample set is relatively small, an increase in the feature space results in poor classification performance, as the classifier based on such training data may be biased and have a large variance.

The curse of dimensionality, often referred to as small sample size (SSS) problem, can be tackled by performing dimensionality reduction of the feature space. Hence, subspace methods, which try to reduce the dimension of the data while retaining the statistical separation property between distinct classes, have been a natural choice for facial feature extraction and recognition. Among the different existing subspace methods, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two powerful tools used for feature extraction and data reduction.

In face recognition, the performance degradation of classical subspace methods is due to the high dimensional face images with larger within class variations than the between-class variations. Therefore, to obtain better performance feature selection, where a subset of features is selected based on a certain criterion, is critical in subspace methods like principal component analysis (PCA) and linear discriminant analysis (LDA). However, the problem of finding a subset of features from a high dimensional feature set is $NP$-hard. Therefore, to solve the feature selection problem, heuristic methods such as evolutionary algorithms are gaining importance. In many face recognition applications, only a small number of training samples for each subject are available. With the small number of training samples, it is difficult to capture all the facial appearance variations due to varying environmental lighting conditions and different facial expressions. In such cases, a classifier, constructed by using a subset of features obtained from small training sets is biased and has a large variance. Consequently, such a classifier may have a poor performance. In classification, the poor performance of a single classifier is tackled by constructing many weak classifiers instead of a single classifier and combines them in some way to obtain a powerful decision rule. Recently a number of such combining techniques, known as ensemble approaches, have been developed. The most popular ones are bagging [1], boosting [2] and the random subspace method [3].

In this paper, we first present a differential evolution (DE) based PCA feature selection, referred to as FS-DE, by
maximizing the distance between all the classes of the training data for face recognition. The selected feature subset is used to train a classifier for the recognition of test images. Later we present an ensemble-based face recognition approach (En-FR) by using different combinations of feature vectors obtained by maximizing the distance between a subset of the classes. The different subsets of the classes are obtained by bagging [1]. Each subset of the PCA features selected is used to train a classifier. The outputs of the different classifiers are combined using a simple majority voting. In the present work, we employ a nearest neighbor classifier.

The rest of the paper is organized as follows. Section II presents a brief literature review of the two subspace methods, PCA and Fisher’s LDA, and some ensemble approaches in face recognition. Section III presents a brief overview of the DE algorithm. Section IV describes the DE based discriminant PCA feature selection (FS-DE) for face recognition. Section V presents the ensemble face recognition (En-FR) using discriminant PCA features. Section VI presents the experimental results while Section VII concludes the paper.

II. ENSEMBLE LEARNING FOR FACE RECOGNITION

PCA [4], as an unsupervised method, is designed to perform dimensionality reduction by maximizing the variance over the entire training samples. Given \( m \) training images, each represented as a vector of \( n \) pixels; the mean image of the training samples is computed and subtracted from each training image, resulting in “centered” images. A matrix \( X \) is formed by collecting all the centered images. The covariance matrix \( \Omega = XX^T \) characterizes the distribution of the \( m \) training images in \( \mathbb{R}^n \). A subset of the eigenvectors of \( \Omega \) is selected. The eigenvectors of the selected subset form a subspace in which the training and the test images are compared. Typically, only \( d \) eigenvectors with the largest eigenvalues are used to define the subspace, where \( d \) is the desired subspace dimensionality.

The traditional motivation to select the eigenvectors with the largest eigenvalues is that the eigenvalues represent the amount of variance along the particular eigenvector. However, in applications like face recognition, selecting the eigenvectors with the maximum variance or higher eigenvalue may not give good recognition accuracy due to their low discriminatory power. In literature, various suggestions have been provided on how to select the eigenvectors for better face recognition [5]. However, the selection of the eigenvectors is problem dependent and requires domain knowledge and expertise.

It is generally believed that the supervised methods that utilize class label information in the training stage, unlike the unsupervised methods, are more effective in dealing with recognition problems. Linear discriminant analysis (LDA) aims to find a set of optimal discriminant vectors by mapping the original data into a low-dimensional feature space. LDA tries to find a set of projection basis by maximizing the between-class scatter matrix and minimizing the within-class scatter matrix in the reduced subspace. In applications like face recognition where the number of training samples is less than the dimensionality of the samples, the scatter matrices degenerate [6] and becomes singular. Fisher faces [7] method mitigates the shortcoming by first performing PCA and then LDA in PCA subspace. In other words, PCA is used as a pre-processing step for dimensionality reduction so as to discard the null space of the within-class scatter matrix of the training data set. Then LDA is performed in the lower dimensional PCA subspace [7]. However, it has been shown [8] that the discarded null space may contain significant discriminatory information [9, 10], which would be beneficial in applications like face recognition.

Therefore feature selection is a critical issue in face recognition systems [8]. Given a set of \( n \) features, the task of selecting a subset of \( d \) features that leads to the smallest generalization error is defined as the feature selection problem. Hence, feature selection reduces the dimensionality of feature space, removes redundant, irrelevant, or noisy data. Finding the best feature subset from a total number of \( 2^n \) candidate subsets for a given \( n \) features can be viewed as a \( NP \)-hard problem. Sequential forward selection (SFS) and sequential backward selection (SBS) are the two well-known heuristic feature selection problems present in the literature [11]. The SFS starts with an empty set and selects the best single feature and then adds that feature to the feature set. The SBS starts with the entire feature set and at each step drops the feature whose absence least decreases the performance. However, these strategies make local decisions and are not expected to find global optimal solutions. Recently, feature selection using heuristic search methods such as evolutionary algorithms is gaining importance [12-16], due to their ability to successfully handle complex optimization problems.

In unconstrained environments, the face recognition systems encounter difficulty in handling the variations in expression, pose and illumination with the limited number of training samples available. In other words, the available training samples may not be representative to handle the different complicated facial variations that arrive during testing. Therefore, subspace face recognition often suffers from the problems such as: 1) small training sample set compared to the high dimensionality feature vector; 2) the performance is sensitive to the subspace dimension. Therefore, instead of pursuing a single optimal subspace, an ensemble learning framework based on different subspaces can be promising.

The most popular ensemble approaches in machine learning are bagging [1], boosting [2] and the random subspace method (RSM) [3]. Bagging is an ensemble
A technique based on bootstrapping and aggregating concepts. Bootstrapping is a technique for random sampling of a new learning (sub) set with replacement from the initial learning data set. Hence, in bootstrapping each example can be selected more than once. Aggregating is a way of combining the outputs of different models. Therefore, in bagging, a collection of base models are learned on different sampled sets generated with bootstrap sampling and their outputs are combined by aggregating techniques such as majority voting. Such an ensemble often gives better results than its individual base models because it combines the advantages of the individual models [7]. Boosting [2], similar to bagging, combine the predictions of base models learned by a single learning algorithm. However, the difference between the two approaches is that in bagging the complementarity of the constructed base models is left to chance, while in boosting we try to generate complementary base models by learning subsequent models, taking into account the mistakes of previous models. In boosting, classifiers and training sets are obtained in a strictly deterministic way. Both training data sets and classifiers are obtained sequentially in the algorithm, in contrast to bagging, where training sets and classifiers are obtained randomly and independently (parallel) from the previous step of the algorithm. In the random subspace method, classifiers are constructed in random subspaces of the data feature space. These classifiers are usually combined by simple majority voting in the final decision rule. With growing attention to ensemble learning, in recent years various ensemble methods, bagging [17, 18], boosting [19-21] and RSM [22-25], for face recognition have been proposed.

III. DIFFERENTIAL EVOLUTION ALGORITHM

Differential Evolution (DE) [26, 27] is as a simple and reliable optimization technique with less number of parameters to tune. Being a parallel direct search method, DE starts by randomly initializing $NP$ $D$-dimensional parameter vectors, $x_{i,G} = \{v_{1,i,G},...,v_{D,i,G}\}$, $i = 1,...,NP$ in the search space constrained by the minimum and maximum parameter bounds $x_{\min} = \{x_{\min,1},...,x_{\min,D}\}$ and $x_{\max} = \{x_{\max,1},...,x_{\max,D}\}$. $G$ is the generation count. In DE, the population evolves over generations by mutation, crossover and selection.

In classical DE, after initialization in generation $G = 0$, successive populations are generated by adding the weighted difference of two randomly selected vectors to a third randomly selected vector, referred to as mutation. By mutation, for each target vector $x_{i,G}$ at generation $G$, its associated mutant vector $V_{i,G} = \{v_{1,i,G},...,v_{D,i,G}\}$ can be generated as

"DE/rand/1": 

$$V_{i,G} = X_{r_1,i,G} + F(X_{r_2,i,G} - X_{r_3,i,G})$$

(1)

The indices $r_1$, $r_2$, $r_3$, $r'_1$, $r'_2$ are mutually exclusive integers randomly generated within the range $[1, NP]$, which are also different from the index $i$. These indices are randomly generated once for each mutant vector. The scale factor $F$ is a positive control parameter for scaling the difference vector. The scale factor $F$ is a user-specified within the range $[0, 1]$.

After mutation, crossover operation is applied to each pair of the target vector $x_{i,G}$ and its corresponding mutant vector $V_{i,G}$ to generate a trial vector: $U_{i,G} = \{u_{1,i,G},...,u_{D,i,G}\}$. In the basic version, DE employs the binomial (uniform) crossover defined as:

$$u_{j,i,G} = \begin{cases} v_{j,i,G} & \text{if } (rand[j] \leq CR) \text{ or } (j = j_{\text{rand}}) \\ x_{j,i,G} & \text{otherwise} \end{cases}$$

(2)

The crossover rate $CR$ is a user-specified constant within the range $[0, 1]$, which controls the fraction of parameter values copied from the mutant vector. $j_{\text{rand}}$ is a randomly chosen integer in the range $[1, D]$.

Then the objective function values of all trial vectors are evaluated and a selection operation is performed as:

$$X_{i,G+1} = \begin{cases} U_{i,G} & \text{if } U_{i,G} \text{ is superior to } X_{i,G} \\ X_{i,G} & \text{otherwise} \end{cases}$$

(3)

The mutation, crossover and selection operations are repeated generation after generation until a termination criterion is satisfied. In the present study, the termination criterion is the maximum number of function evaluations used by the algorithm. The algorithmic description of the DE is summarized in Table 1.

| Step 1 | Set the generation count $G = 0$. Randomly initialize a population of $NP$ individuals $X_{i,G} = \{v_{1,i,G},...,v_{D,i,G}\}$, $i = 1,...,NP$ uniformly distributed in the search range. Evaluate the individuals based on the objective function provided. |
| Step 2 | WHILE stopping criterion is not satisfied DO FOR $i = 1$ to $NP$ |
| Step 2.1 Mutation step | Generate a mutated vector $v_{i,G} = \{v_{1,i,G},...,v_{D,i,G}\}$ corresponding to the target vector $X_{i,G}$ using the mutation strategy given by the eqn. (1). |
| Step 2.2 Crossover step | Generate a trial vector $U_{i,G} = \{u_{1,i,G},...,u_{D,i,G}\}$ for each target vector $X_{i,G}$ through binomial crossover given by the eqn.(2) |
| Step 2.3 Selection step | Evaluate the trial vector $U_{i,G}$ and the population for the next generation is updated using the selection step given by eqn. (3) END FOR |
| Step 2.4 Increment the generation count $G = G + 1$ END WHILE |
IV. DISCRIMINANT PCA FEATURE SELECTION USING DIFFERENTIAL EVOLUTION ALGORITHM

In the present work, the problem of discriminant feature subset selection from a set of PCA features is done by using differential evolution (DE) algorithm presented in Section III. In other words, the aim of the algorithm is to select $d$ discriminant features from a set of $n$ PCA features. In DE, each population member encodes the indices of the $d$ features to be selected from the $n$ features. In feature selection, the indices of the selected features are integer variables and DE can handle only continuous variables. To handle the integer variables of the feature selection problem certain modifications are incorporated in the DE algorithm. Even though, the population of DE encodes real values and evolve through generations, during the evaluation, the parameters of the population members are converted to integers by using the rounding operation.

The fitness function that guides the entire search process of the DE algorithm to find a best subset of discriminant PCA features is defined as:

$$ f(X) = \arg \max_{X} \left| X S_{B} X^{T} \right| - 100 \cdot N^{2} \quad (4) $$

In equation (4), the first term on the right side maximizes the distance between the classes ($S_{B}$) of the training data, where

$$ S_{B} = \sum_{i=1}^{C} n_{i}(\mu_{i} - \mu)(\mu_{i} - \mu)^{T} \quad (5) $$

$X$ contains the eigenvectors or features corresponding to the $d$ indices encoded in the individual population members. $n_{i}$ is the number of elements in class $i$, $\mu$ is the mean of the images in entire training data set. $\mu_{i}$ is the mean of the images present in class $i$ of the training data set.

The second term in equation (4) prevents the selection of the same PCA features by penalizing the individual population members. $N$ represents the number of repeated indices of the feature vectors.

In this paper, we try to obtain a subset of discriminant PCA features for face recognition using the DE algorithm. The DE algorithm selects the discriminant PCA features by maximizing the objective given by equation (4).

V. ENSEMBLE BASED FACE RECOGNITION USING DISCRIMINANT PCA FEATURES

A single subset of PCA features obtained by maximizing the class separation may not properly separate all the classes in the training data. Therefore, the classifier can be biased and may not provide good classification accuracy. In other words, the separability criteria may not be directly related to the classification accuracy in the output space. Based on the ensemble learning, it would be better to use a combination of different subsets of PCA features and combine their outputs.

In this section, we present an ensemble based face recognition algorithm using discriminant PCA features. Using bagging, different subsets of the classes are sampled from the original set of the entire classes. Due to the bootstrap sampling, the subsets obtained overlap each other. Corresponding to each subset of the classes, a PCA feature subset is obtained by using the DE algorithm presented in Section IV, by maximizing the distance between the classes in the subset. The different PCA feature subsets obtained can be used to classify the test images and the outputs are combined by majority voting.

In En-FR, as the number of classes in the subset is small, a better class separability can be obtaining by maximizing the distance between the classes in the subset. Therefore, the subset of feature vectors can better classify the classes in the particular subset. As the subset of classes obtained by bagging have an overlap between them and the prediction of all the classifiers are combined by majority voting, the prediction accuracy of the ensemble approach can be better and robust.

VI. EXPERIMENTAL SETUP AND RESULTS

To evaluate the algorithms four well-known face image databases are considered. The details of the databases are summarized in Table 2 and the samples corresponding to different databases are presented in Figs 1-4. On every database, each algorithm is repeated 30 times. During each run, on a database, 3 randomly selected samples per class are used for training, while the remaining samples are used for testing. All the images are resized to 32x32 pixels. Therefore, the number of original PCA features are $n = 1024$. 

![Figure 1. Samples images of Yale A Face Database](image1.png)  
![Figure 2. Samples images of ORL Face Database](image2.png)
Table 2. Properties of Different Databases

<table>
<thead>
<tr>
<th>Database</th>
<th>No. of Classes</th>
<th>No. of Samples per Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yale A [28]</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>ORL [29]</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>AR [30]</td>
<td>50</td>
<td>13</td>
</tr>
<tr>
<td>Yale B [31]</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

A. Experiment 1:
The aim of the experiment is to show that the selection of PCA eigenvectors based on the eigenvalues is not always good for face recognition and the selection of different combination of eigenvectors is better for different databases. In the present experiment, the PCA eigenvectors obtained from the training data are sorted in the decreasing order of magnitude of their eigenvalues. Different combinations of the eigenvectors are selected and are used to face recognition. The different combinations of the eigenvectors used are:

S1: Select five eigenvectors with indices one to five
S2: Select five eigenvectors with indices two to six
S3: Select five eigenvectors with indices three to seven

The average recognition accuracies of the different combinations of the PCA eigenvectors are presented in Table 3.

Table 3. Recognition Accuracies of Different Feature Selections

<table>
<thead>
<tr>
<th>Database</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yale A</td>
<td>63.06%</td>
<td>57.39%</td>
<td>53.86%</td>
</tr>
<tr>
<td>ORL</td>
<td>62.93%</td>
<td>60.37%</td>
<td>54.79%</td>
</tr>
<tr>
<td>AR</td>
<td>18.92%</td>
<td>20.33%</td>
<td>21.67%</td>
</tr>
<tr>
<td>Yale B</td>
<td>76.47%</td>
<td>77.18%</td>
<td>77.45%</td>
</tr>
</tbody>
</table>

From the results, it can be observed that the S1 combination of features give good face recognition accuracies for Yale A and ORL databases. For AR and Yale B databases, the S3 combination of features gives better recognition accuracies. The S2 combination of features is better than S3 for Yale A and ORL databases and gives similar performance on AR and Yale B datasets. However, the performance of S2 combination is better than S1 on AR and Yale B datasets while the performance on Yale A and ORL is worse than S1. From the results, it can be observed that, depending on the database different combinations of PCA vectors can better. We also observe that, in a given database, the performance of different combinations vary depending on the training samples selected. This may be due to the fact that the eigenvectors selected based on maximizing the variation does not provide good class separability and are not good for face recognition. Therefore, from the results it can be observed that the selection of eigenvectors for face recognition is not straightforward and requires trial-and-error search.

B. Experiment 2:
To demonstrate the superior performance of DE based PCA feature selection algorithm (FS-DE) over the Fisher LDA based algorithm and selection of PCA eigenvectors based on the magnitude of the eigenvalues. The Fisher LDA algorithm operates on the first five features of the PCA subspace. Based on the algorithm in Section IV, we select five PCA eigenvectors by maximizing the distance between all the classes using the DE algorithm. The selected eigenvectors along with a nearest neighbor classifier are used to classify the test images.

In the present study the parameters of the DE algorithm are set as: $NP = 50$, Maximum Number of Function Evaluations $= 50,000$, $CR = 0.3$, $F = 0.5$. The maximum and minimum bounds for the parameters are $1$ and $1024$, respectively. The average recognition accuracies of the algorithms (Fisher LDA and FS-DE) are summarized in the first two columns of Table 4.

Comparing the results in Tables 3 & 4, it can be observed that the performance of the FS-DE algorithm is better than the different combinations of the PCA feature vectors. The better performance of FS-DE compared to the different combinations of the eigenvectors can be attributed to the DE selection of the eigenvectors by maximizing the class separation and not based on maximizing the variance.
From the results in Table 4, it can be observed that the DE based selection of the PCA feature subset gives similar or better recognition accuracies on different face databases. The proposed FS-DE algorithm shows superior performance compared to Fisher LDA algorithm on the AR database. Both Fisher LDA and FS-DE try to maximize the class separation. However, the performance of FS-DE is better than Fisher LDA because Fisher LDA operates on the PCA subspace in which the eigenvectors are selected based on maximizing the variation and not the class separability. The discarded subspace of PCA with minimum variance may contain better discriminatory information. In FS-DE, the eigenvectors are selected based on maximizing the class separation.

C. Experiment 4:

The aim of the experiment is to show that the proposed En-FR method is better than the single feature subset selection. Using bagging, 20 subsets of the classes are selected. Each subset contains \((1/5)\)th of the total number of classes in the training dataset. Obtain five PCA eigenvectors, corresponding to every subset of classes, by maximizing the distance between the classes in the subset, using DE algorithm. Each feature subset along with a nearest neighbor is used to classify the test images. The outputs of the different classifiers are combined by a simple majority voting.

The average accuracy results of the proposed En-FR algorithm on different databases are summarized in the third column of Table 4.

<table>
<thead>
<tr>
<th>Database</th>
<th>Fisher LDA</th>
<th>FS-DE</th>
<th>En-FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yale A</td>
<td>61.81%</td>
<td>63.72%</td>
<td>69.53%</td>
</tr>
<tr>
<td>ORL</td>
<td>64.61%</td>
<td>63.12%</td>
<td>66.26%</td>
</tr>
<tr>
<td>AR</td>
<td>18.47%</td>
<td>22.51%</td>
<td>26.71%</td>
</tr>
<tr>
<td>Yale B</td>
<td>79.43%</td>
<td>78.12%</td>
<td>80.37%</td>
</tr>
</tbody>
</table>

From the results, it can be observed that the performance of the ensemble is superior to a single subset PCA feature selection. In En-FR, each subset PCA features maximize the class separation of a subset of classes instead of the entire classes in the training data. Hence, the class separation by the different subset feature vectors in En-FR is better than a single feature subset selection. Therefore, the better performance of En-FR is a result of better class separation due to the ensemble approach. The En-FR algorithm is robust and provides better generalization ability.

In the present study, the ensemble approach is carried out by obtaining 20 bagging samples of the total classes and each contains \((1/5)\)th of the total number of classes in the training set. The computational times taken by the Fisher LDA, FS-DE and En-FR algorithms are 2.4, 6.4 and 80 seconds respectively. From the experiments, we have observed that the performance of the En-FR algorithm improve with the increase in the number of bagging samples. The recognition accuracies of the proposed algorithm on Yale A, ORL, AR and Yale B databases when the number of bagging samples is increased to 30 is 70.12, 66.93, 27.10 and 81.20 respectively. However, increasing the number of bagging samples increases the computational load of the algorithm.

VII. CONCLUSIONS

From the experimental results, it has been demonstrated that the selection of PCA eigenvectors based on maximizing the variance is not apt for face recognition. The selection of a subset of PCA eigenvectors for face recognition is problem dependent and is not straight forward. It can also be observed that the LDA is not always superior to PCA in face recognition applications. Therefore, to obtain better face recognition, feature selection is necessary in face recognition applications.

In this paper, we propose a DE based PCA feature selection to solve the \(NP\)-hard problem of feature selection in face recognition. From the experimental results, it can be observed that the DE based feature selection (FS-DE) is superior to the random selection of the PCA feature vectors. Based on the intuition, that a combination of many classifiers perform better than a single classifier, we propose and ensemble based face recognition algorithm (En-FR). The experimental results demonstrate the superior performance of the ensemble approach.

ACKNOWLEDGMENT

This research was supported by the Converging Research Center Program funded by the Ministry of Education, Science and Technology (2011K000659), Ministry of Education, Science Technology (MEST) and Korea Institute for Advancement of Technology(KIAT) through the Human Resource Training Project for Regional Innovation

REFERENCES


